

Responses to Standards Question

I - Metadata Standards

1. Metadata standards, particularly international standards such as ISO 19115. Metadata is essential from a user perspective, because without it, dataset quality and suitability for purpose cannot be ascertained. However, there are currently many standards in use, and many users ignore standards completely. Many data managers are against the idea of using a single metadata standard, often because they can be restrictive in terms of the fields and concepts they describe. Many are therefore tempted to abandon recognized data standards in favour of their own systems that are viewed as being more descriptive of specific data sets. However, metadata standards are needed in order to allow catalogue interoperability. Without the use of agreed standards by all participants, any GMES system or service infrastructure will be reduced to a data depository.
2. Use of XML. The use of XML is already becoming an accepted standard in many applications, not through any particular agreement, but because of the widespread recognition of the benefits it offers. XML is a particularly useful standard because of its ability to handle complex data sets and provide flexible and adaptable information identification. An initiative such as GMES is likely to involve the transfer of a variety of complex data sets from a variety of sources over the web, and therefore requires a flexible standard such as XML.
3. Metadata content (to facilitate semantic and other translations required for content interoperability).

II - Data Access Standards

1. OPeNDAP is the suggested middleware for the Integrated Ocean Observing System, and as we are working with several projects funded as part of the IOOS, it is something that is likely to grow in importance in our data handling.
2. OPeNDAP protocol.
3. The E-Government Act of 2002 calls for the U.S. Federal Government to enhance interoperability by adopting a common search standard. Now posted for public comment is a draft "Statement of Requirements for Search Interoperability" at <http://www.search.gov/interop/requirements.html> (Actual experiences with standards-based search interoperability will be of particular interest.)

This draft is based on a recommendation of the OASIS E-Government Technical Committee, and aligns with findings of a white paper by the Industry Advisory Council [IAC], Enterprise Architecture Shared Interest Group, titled "Interoperability Strategy: Concepts, Challenges, and Recommendations".

4. I would be very interested in having your standards effort consider the OPeNDAP data access protocol. Not sure how much you know about it. The idea is to facilitate access to data over the network. The protocol is discipline neutral, that is a standard for the semantics associated with a data set are not imposed on the data provider. This means that the protocol may be used by virtually any scientific (or other for that matter) discipline. The way that we had envisioned it being used (and the way that it is in fact used) is for different disciplines to layer the semantics on top of the protocol. The protocol effectively isolates the user from the format in which the data are stored. It is in heavy use in the meteorological and oceanographic communities - > 500 data sets served at present with more than 2000 active users per month making more than 2.5 million accesses. It is also being used in some areas of space physics.

We are currently modifying the protocol to operate in a web services environment. We refer to this as DAP4. We anticipate the first implementation of DAP4 will be available this summer.

As noted above the OPeNDAP data access protocol addresses format dependence. We are now working on the layers above this, layers associated with the semantics of the data for oceanographic applications and the restructuring of what we refer to as sequence data, generally in situ data.

III - Data Order Notification Standards

1. A standard means of notification for asynchronous (batch) order fulfillment could extend the reach of such things as automated data mining, applications, etc. beyond online disk and into the tape archives of EOSDIS.
One area where a standard could enable something that is difficult today is the area of order fulfillment notifications. Data orders imply an asynchronous process, such as the retrieval of data from tape and staging to disk, so that the typical scenario is:
 - a. client requests data or service
 - b. provider accepts request
 - c. provider spends some time (minutes to days) filling request
 - d. AT SOME LATER TIME, provider notifies client that output is ready for pickup

The means of notification in Step 4 has no standard: notifications are currently via either email or FTP-push. The formats of the notifications are unique to each data or service provider. This makes it difficult to construct clients (esp. in the field of applications) with fully automated machine-to-machine interfaces: that is, an interface that could request data (or services) and when the data are ready, automatically retrieve it.

IV - Data Format Standards

1. We suggest that ESML be adopted as an ESE standard, to supplement the variety of data formats already in use.

From our experience working with Earth Science data users and data systems, we have found that:

- a. One of the most time-consuming aspects of using a new data product is initially deciphering what's in the data and how to read it into a model or analysis tool
- b. Nevertheless, neither data producers nor data users are particularly excited about "standard data formats" and one standard format doesn't fit all data products.

ESML was developed in response to this situation. Based on XML, ESML was designed as an elegant solution to address the research issue of data format heterogeneity, and to specifically target the special characteristics of science data and spatial imagery. ESML is unique in that it is not another new data format; instead it is an external syntactic metadata based solution for decoding existing formats. ESML provides a language for representing scientific data formats and structures, and its associated software library enables integration of disparate and often distributed data.

2. The two most needed standards for validation datasets (from ground based, aircraft and satellite) are HDF4/5 and the Envisat/Aura Validation Program metadata guidelines for correlative data.
3. Gridding schemes (to facilitate data fusion).
4. Defining a set of standard formats used for data distribution is viewed as the primary area that would be beneficial. These formats should be compatible with commercially available analysis tools. Standards for file naming that are self explanatory and include a form of data location and acquisition time information, and for parameter naming of the primary resultant fields, would also be established.

The standard formats would constitute a distribution standard, not a processing or archive standard. This implies a flexible generation capability that is aligned with a "backend" distribution function where data is placed in some requested standard form for delivery to the user.

In conjunction with this would be standard access capabilities supplied through generally available I/O libraries to ease software development efforts.

5. We have been using the Point Structure of HDF-EOS as a way to store and sort through observation data. The HDF-EOS support for the Point Structure has been minimal - one or two what we consider essential calls were dropped in the HDF-EOS 5 version (there was a call that gave you all the records from a higher level (in number - further down the tree) that correspond to a value of the link at a lower level (in number) that was dropped), plus I am not certain that the implementation of the Point Structure takes advantage of features in HDF5 that weren't available in HDF4. As an example of what might make good use of HDF5 for observational data is pytables (<http://pytables.sourceforge.net/html/WelcomePage.html>), I don't know if you have ever looked at it but it has many of the really desirable features for dealing with observational data, the drawback being that you have to use Python. It is, however, built on top of HDF5, and shows much more thought about using the features of HDF5 than the HDF-EOS 5 Point Structure.

The other problem is that as compared to gridded data, where it is reaching a point where there are only a handful of formats being used, there is almost no standardization for observational data.

Secondly, the issue of standardizing what is put into a data format. For example, the fact that you have dimension or units in NetCDF doesn't tell you what they are, and for a program to use these things effectively, these elements must also be standardized. For our netcdf gridded files we pretty much use the COARDS convention, though many people use the expanded CF convention. Having these, or something similar, as part of the standard we also would consider essential. Also, to beat a dead horse, there is no similar set of standards for observational data. The best I have seen so far are those for the ARGO program, using NetCDF files - but the contents, units etc etc are very precisely laid out.

6. CF netCDF conventions.
7. Standard imagery formats could significantly reduce the amount of pre-processing required to get imagery into useful formats for analysis. For our work, and relative to the variety of image data sources we work with, the GeoTIFF format is a standard that we would like to see. Imagery sources that we use include: MODIS, Landsat TM, and various commercially available image types.

V - Web Security Standards.

1. Our REASoN research project, "A Border Security Decision Support System Driven by Remotely Sensed Data Inputs," will be involved with some sensitive geospatial data for homeland security tasks. Currently, the NASA ESE does not provide sufficient guidelines or standards regarding the web-based data security and encryption methods. It will be great if the SEEDS SPG can initiate a technical working group focusing on the web-based data/information security standards. The web security standard will focus on the following issues
 - How to protect the data access/download procedures via the Internet?
 - How to limit the access of web mapping servers only to authorized users?
 - How to create different level of access to web services (GIS functions) for different uses?
 - What kinds of encryption methods or security techniques do the ESE recommended?

VI - Others:

1. Spatial Subsetting Capability for NASA EOSDIS.
A spatial subsetting capability is missing in the EOSDIS, which is particularly important for users of land remote sensing data. Users should have an ability to request and download image sets by geographic extent rather than be arbitrary spatial units.
2. STANDARD GLOBAL THREE-DIMENSIONAL SPHERICAL REPRESENTATION OF A DIGITAL EARTH MODEL
The map projection grids currently in use provide for two dimensional representations of the Earth surface. None of these projections is suitable for a global unified interaction with a digital Earth model. For that purpose, a uniform, continuous, conjugate, and global digital expression of the Earth sphere needs to be established.

Currently there are individual approaches being developed that include methodologies very similar to what is required, but such efforts need to be directed in a manner that results in standards that are adopted community-wide and industry-wide. These approaches use terms like "Octahedral Quaternary Triangular Mesh" and "Geodesic Grid" to describe methods for spherical surface tessellation based on polyhedrons.

3. STANDARD ATTITUDE AND EPHEMERIS

Continue the current EOS mission concept of a standard set of attitude and ephemeris data measurements and file structures. Also continue providing standard access and transformation libraries for these data sets, as is currently available in the EOS Toolkit.

4. Our typical example is the ability to discover data without knowing a file name, access the data without knowing where it is stored, retrieve the data without knowing what access protocol is needed by the remote storage system, and manipulate the data through a preferred interface.

We work with communities that build digital libraries for publishing data, data grids for sharing data, and persistent archives for preserving data. We use generic data management infrastructure (Storage Resource Broker) to implement all three environments. Based on interactions with these communities, the standard approach is to assemble a collection of data that is going to be shared. The collection can span multiple sites.

To manage data distributed across multiple sites, we had to implement virtualization mechanisms (you need at least 7). They first four are

- storage repository virtualization. This is the set of operations one uses to interact with a remote storage system. See the paper in the Global Grid Forum on "Operations for Access, Management, and Transport at Remote Sites"
- data virtualization. This is the provision of a naming convention for distributed data that is infrastructure independent. The context for management and discovery of the distributed digital entities is mapped onto the logical name space. Types of attributes include administrative (where the data is located, owner, access controls, size, etc.), descriptive (metadata attributes used for discovery, provenance), authenticity (checksums, audit trails).
- information storage repository virtualization. The context for the distributed data is stored in a database as a catalog. The operations used to manage the catalog can be characterized (bulk metadata load, schema extension, automated SQL generation, export and import of XML files, etc.). This makes it possible to store the context in any vendor database.
- access virtualization. This is the set of operations that can be performed by users on the combined data and information storage repositories. The standard set is mapped to the particular API desired by the user (Perl, Python, WSDL, Web browser, Java, C++, OAI, etc.).

The result is the ability to organize distributed data into a collection, share data between sites and users under access controls, discover data based on metadata attributes, track all operations done on data.

The closest to your scenario is the sharing of data between data grids. This requires controls on sharing of resources (may I write on your storage), user names (who authenticates your name), files (which files are shared for whom), and metadata (who updates metadata to track operations on the files). Interchanges between data grids requires specifying all of these constraints.

5. Develop a flexible system at the least. We have one slow, outdated, expensive to maintain SGI computer to run the EOSHDF Toolkit for our EOS instrument. Everything else we have has migrated to LINUX PCs with 10 times the performance with a 3yr warranty and 3 TB of storage, and the cost of each system is the same as 1 yr of maintenance on our old SGI.

General Comments

1. I am writing to advise you of an NSF-funded activity that is looking at exactly these issues. We have reports and notes on these topics on our web-site <http://bbe.sdsc.edu/InteropWorkshop> . These are still in draft form for the most part but they are fairly extensive.
2. A group at GSFC last year asked for similar information based on lessons learned on-center. I don't recall details, but I'm sure someone else here knows further details. They asked for white papers of ~1-2 pp from each volunteer contributor. Fred Huemrich might have more information as he led a one of the responses.

The CEOS WGISS group is actively attempting to develop a test system where data from different sources are projected into a common format, etc., upon user request. Any/all reformatting would occur transparently to the user (i.e., behind the scenes). POC is: "Tim B Smith" . This project is in cooperation with CEOS WGCV, which is currently headed by a GSFC person (Steve Ungar). Jeff Morissette is the local NASA POC for the test system.

3. Please find attached a general information document describing FAO's spatial information infrastructure development under GeoNetwork. Furthermore I attach a recent FAO status report to the UN Geographic Information Working Group (UNGIWG) which has more detailed information on FAO activities related to the use of spatial information.